Speaker age estimation using i-vectors

Mohamad Hasan Bahari[a,*], Mitchell McLaren[b], Hugo Van hamme[a], David A. van Leeuwen[b]

[a] *Center for processing speech and images, KU Leuven, Belgium*
[b] *Center for Language and Speech Technology, Radboud University Nijmegen, The Netherlands*

Abstract

In this paper, a new approach for age estimation from speech signals based on i-vectors is proposed. In this method, each utterance is modeled by its corresponding i-vector. Then, a Within-Class Covariance Normalization technique is used for session variability compensation. Finally, a least squares support vector regression (LSSVR) is applied to estimate the age of speakers. The proposed method is trained and tested on telephone conversations of the National Institute for Standard and Technology (NIST) 2010 and 2008 speaker recognition evaluation databases. Evaluation results show that the proposed method yields significantly lower mean absolute error and higher Pearson correlation coefficient between chronological speaker age and estimated speaker age compared to different conventional schemes. The obtained relative improvements of mean absolute error and correlation coefficient compared to our best baseline system are around 5% and 2% respectively. Finally, the

*Corresponding author. Tel:+32-(0)16-32.85.45. Fax:+32-(0)16-32.17.23.
Email addresses: mohamadhasan.bahari@esat.kuleuven.be (Mohamad Hasan Bahari), m.mclaren@let.ru.nl (Mitchell McLaren), hugo.vanhamme@esat.kuleuven.be (Hugo Van hamme), d.vanleeuwen@let.ru.nl (David A. van Leeuwen)

effect of some major factors influencing the proposed age estimation system, namely utterance length and spoken language are analyzed.
Keywords: speaker age estimation, i-vector, least squares support vector regression, utterance length, language mismatch.

1. Introduction

Automatic identification of age from speech signals has a wide range of commercial and forensic applications (Dobry et al., 2011; Tanner and Tanner, 2004; Li et al., 2013). For example, in targeted advertising through internet, where user-computer and user-company vocal interaction has increased significantly during the last decades, information about the user's language/accent, age and gender can help to offer appropriate products and services (Schuller et al., 2013). Speaker age estimation is also required in many forensic scenarios such as kidnapping, threatening calls and false alarms to facilitate the identification of criminals, e.g. to narrow down the number of suspects (Tanner and Tanner, 2004). This technology can also guide ambient assisted living and smart home systems to automatically adapt to different user needs (Li et al., 2013). Speaker age estimation is also required in natural human-machine interaction. In video games, knowledge about the user's age and gender can help to adapt the game accordingly (Schuller et al., 2013). Automatic identification of speaker age can be applied to improve the performance of other speech technology systems such as emotional state recognition, smoker speaker detection, identifying the level of intoxication and even automatic speech recognition (ASR).

Experimental studies reveal major effects of vocal aging on the speech sig-

nal such as lowered speaking rate and increased jitter and shimmer (Schotz, 2006), and has shown to negatively influence speaker recognition performance (Kelly et al., 2013). However, the relation of these acoustic cues with speaker age is usually complex and affected by many other factors such as speech content, language, gender, weight, height, emotional condition, smoking and drinking habits (Schotz, 2006; Bahari and Van hamme, 2011; Bahari et al., 2012b). Furthermore, in many practical cases we have no control over the available speech duration, content, language, etc.. These issues make automatic speaker age estimation very challenging for both humans and machines (Bocklet et al., 2008b; Li et al., 2013; Schotz, 2006).

Figure 1, which shows a simplified model for human speech production, helps to display the underling difficulties in speaker age estimation. In this problem, the recorded speech signal is the only available information and the task is to estimate one of the physical states of the articulatory system, namely the speaker's age, without any information about the system inputs, channel characteristics and the other psychological and physical states of the articulatory system such as gender, emotional state and smoking habit.

Technical factors such as available speech duration, environment, recording device and channel conditions also influence the estimation accuracy. In other words, in a typical practical scenario, the quality of the available speech

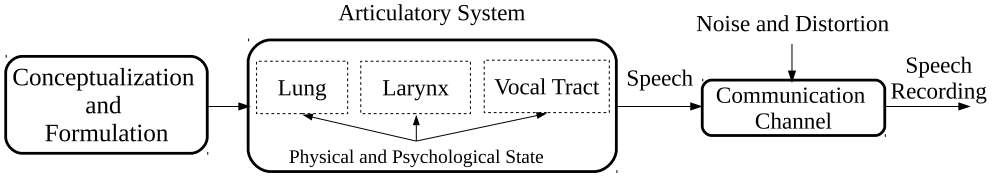

Figure 1: *simplified human speech production model and recording channel.*

signal and the recording conditions are not controlled and the duration of the speech signal may vary from a few seconds to several hours.

1.1. Related Work

Studies on the influence of ageing on voice started in the late 1950s (Mysak, 1959). However, the first automatic speaker age recognition systems were developed around four decades later in the early 2000s (Linville, 2001; Muller et al., 2003; Minematsu et al., 2002; Shafran et al., 2003). During this decade, many different techniques, mostly inspired from the automatic speaker and language recognition fields, have been suggested for categorizing speakers based on their age groups. For example, using different types of acoustic features and Support Vector Machines (SVM) (Mahmoodi et al., 2011; Chen et al., 2011; van Heerden et al., 2010), Gaussian Mixture Model (GMM) mean supervectors and SVM (Bocklet et al., 2008a), nuisance attribute projection (Dobry et al., 2009), anchor models (Dobry et al., 2009) and parallel phoneme recognizers (Metze et al., 2007). The age sub-challenge of the Interspeech 2010 paralinguistic challenge provided a forum for presenting state-of-the-art methods in speaker age group classification (Schuller et al., 2010). Participants of the age sub-challenge tried to categorize speakers of telephony data in the "aGender" corpus into four age groups — 7 to 14 (Child), 15 to 24 (Youth), 25 to 54 (Adult) and 55 to 80 (Senior) years old. In this sub-challenge, GMM mean supervectors (Bocklet et al., 2010), GMM weight supervectors (Porat et al., 2010), Maximum-Mutual-Information (MMI) training (Kockmann et al., 2010) and fuzzy SVM modeling (Nguyen et al., 2010) have been suggested to enhance acoustic modeling quality. A brief overview of different proposed methods in this sub-challenge is presented in (Li et al.,

2013), which also introduces an age group recognition approach using acoustic and prosodic level information fusion.

In speaker age group recognition, crisp borders are assumed between different age groups. For example, in the mentioned age sub-challenge, a speaker with age 54 belongs to the adult group and a 55 year old speaker belongs to the senior category, These two speakers who have only one year of age difference and share many similarities are considered to be from two different categories, while a 80 year old speaker with distant characteristics is in the same category as the 55 year old speaker. This setup causes many problems in training, testing, and performance measurement. To avoid these troubles, recently it has been suggested to use regression for age estimation (Bocklet et al., 2008b; Dobry et al., 2011; Bahari and Van hamme, 2011; Bahari et al., 2012b; Feld et al., 2009). A probabilistic interpretation of the posterior distribution of age estimation and its calibration is presented in (van Leeuwen and Bahari, 2012).

1.2. Motivations, Goals and Summary of Contributions

One effective approach to age estimation from speech involves modeling speech recordings with Gaussian Mixture Model (GMM) mean supervectors to use them as features in Support Vector Regression (SVR) (Dobry et al., 2011; Bocklet et al., 2008b). Similar Support Vector Machine (SVM) techniques have been successfully applied to different speech processing tasks such as speaker recognition (Campbell et al., 2006). While effective, GMM mean supervectors are of a high dimensionality resulting in high computational cost and difficulty in obtaining a robust model in the context of limited data. Consequently, dimension reduction through PCA-based methods has

been found to improve performance in age estimation from GMM mean supervectors (Dobry et al., 2011). In the field of speaker recognition, recent advances using so-called i-vectors (Dehak et al., 2011a) have increased the classification accuracy considerably. An i-vector is a compact representation of an utterance in the form of a low-dimensional feature vector. The same idea was also applied in speaker or language and dialect recognition effectively (Dehak et al., 2011b; Bahari et al., 2013). In our last paper, we successfully replaced GMM mean supervectors by low-dimensional i-vectors to model utterances in an SVR based speaker age estimation system (Bahari et al., 2012a). The results of evaluation on the NIST 2010 and 2008 SRE databases illustrated that the i-vector based speaker age estimator increases the estimation accuracy.

In this paper, we extended our previous work (Bahari et al., 2012a) by:

1. Applying Within Class Covariance Normalization (WCCN) (Hatch et al., 2006) for session variability compensation. In our last paper (Bahari et al., 2012a), we have applied WCCN to normalize utterances of each age group. This method was not successful. In this paper we updated our strategy to use WCCN, where the classes are speakers rather than age groups.

2. Replacement of SVR by least squares SVR (LSSVR) to improve the computational cost.

3. Updating the evaluation setup to increase the size of training dataset, which helps the classifier to observe more variability in the data.

4. Using a standard z-test to analyze the statistical significance of the obtained improvements by the proposed method.

5. Investigate the effect of utterance length on the proposed automatic

speaker age estimation system.

6. Investigate the language mismatch on the proposed method.

The rest of this paper is organized as follows. In Section 2 the problem of speaker age estimation and different conventional approaches addressing this issue are described. In section 3, the proposed approach is elaborated. Section 4 explains our experimental setup. The evaluation results and an investigation of parameters affecting speaker age estimation are presented and discussed in section 5. The paper ends with conclusions in section 6.

2. Age Estimation from Speech

In speaker age estimation, we are given a training dataset of speech recordings $S^{\text{tr}} = \{(\mathcal{X}_1, y_1), \ldots, (\mathcal{X}_s, y_s), \ldots, (\mathcal{X}_S, y_S)\}$. In this set, $\mathcal{X}_s$ and y_s denote the s^{th} utterance of the training dataset and its corresponding speaker age, respectively. The goal is to design an estimator function $\mathcal{G}$, such that for an utterance of an unseen speaker $\mathcal{X}^{\text{tst}}$, the actual speaker age is predicted accurately.

2.1. Baseline Approaches

In this paper, we use three baseline approaches with which we compare our proposed regression techniques:

Prior: The most basic choice for the estimator function is the average age of the training data, $g(\mathcal{X}^{\text{tst}}) = \frac{1}{S}\sum_s y_s$. This estimator, labeled as *prior* in the rest of this paper, intuitively provides a reference level of accuracy.

GMM-R: Different methods have been introduced to reach an effective speaker age estimation (Dobry et al., 2011)–(Bahari and Van hamme, 2011).

For example, Bocklet et al. introduced GMM-R to estimate the age of children from GMM mean supervectors derived from their utterances (Bocklet et al., 2008b). Given an utterance, Maximum A Posteriori adaptation (MAP) is applied to adapt a Universal Background Model (UBM) to the speech characteristics of the speaker (Campbell et al., 2006). The component means of the obtained GMM are then extracted and concatenated to form a GMM mean supervector representing the utterance. Finally, an SVR is applied as a function approximator to estimate the speakers' age.

GMM-PCA-R and **GMM-WPPCA-R:** The approach of GMM-R was adopted and extended by Dobry et al. (Dobry et al., 2011) by applying dimension reduction techniques to the supervector. Methods such as Principal Component Analysis (PCA) and Weighted-Pairwise PCA (WPPCA) were applied and investigated. It was concluded that WPPCA, which is a supervised dimensionality reduction approach based on nuisance attribute projection (Dobry et al., 2011), yields more accurate results. These speaker age estimators, labeled GMM-PCA-R and GMM-WPPCA-R, are used as contrastive baseline systems in this paper.

3. Age Estimation using i-vectors

This section briefly describes the main components of the i-vector based age estimation approach, namely SVR and LSSVR, the i-vector framework and WCCN. Then, the proposed method is elaborated and finally the proposed scheme is presented.

3.1. Regression

In this section, SVR and LSSVR are briefly introduced.

3.1.1. Support Vector Regression

Support Vector Regression (SVR) is a function approximation approach developed as a regression version of the widely known classification paradigm, namely Support Vector Machines (SVM) (Lu et al., 2009; Smola and Scholkopf, 2004). While SVMs perform the classification task by determining the maximum margin separation hyperplane between two classes, SVRs carry out the regression task by finding the optimal regression hyperplane in which most of training samples lie within an ϵ-margin around this hyperplane (Smola and Scholkopf, 2004). In a typical regression problem a training dataset $S^{\text{tr}} = \{(a_1, b_1), \ldots, (a_n, b_n), \ldots, (a_N, b_N)\} \subset \mathbb{R}^d \times \mathbb{R}$ is given, where a_n and b_n denote model input and corresponding output of the n^{th} data point respectively. The objective of the regression analysis is to determine a function $f(a)$, so as to predict the desired outputs accurately. In the primal form of SVR the following relation is considered for $f(a)$:

$$f(a) = \mathbf{w}^t \Phi(a) + z \tag{1}$$

where $\Phi(a)$ denotes a mapping function in the feature space, $\mathbf{w}$ is a row vector with the same dimension of $\Phi(a)$, $z \in \mathbb{R}$ is a constant and and t represents the transpose operator. Using Vapnik's ϵ-insensitive loss function the model training—estimation of $\mathbf{w}$ and z—is formulated as to minimize

$$\frac{1}{2} \|\mathbf{w}\|^2 + \lambda \sum_{n=1}^{N} (\xi_n + \xi_n^*) \tag{2}$$

subject to

$$\begin{cases} b_n - \mathbf{w}^t \Phi(a_n) - z & \leq \epsilon + \xi_n \\ \mathbf{w}^t \Phi(a_n) + z - b_n & \leq \epsilon + \xi_n^* \\ \xi_n, \xi_n^* & \geq 0. \end{cases} , \tag{3}$$

where ξ_n and ξ_n^* are slack variables vanishing during the optimization process, $\epsilon > 0$ controls the ϵ-insensitive zone used for fitting the training data and $\lambda > 0$ determines the trade-off between the atness of $f(a)$ and the cost of tolerating deviations larger than ϵ.

For high dimensional data, this constrained minimization problem can be solved more efficiently by introducing a dual set of variables and solving the following dual optimization problem (Smola and Scholkopf, 2004)

$$\begin{aligned} \max_{\alpha,\alpha^*} & -\frac{1}{2} \sum_{m,n=1}^{N} (\alpha_n - \alpha_n^*)(\alpha_m - \alpha_m^*) \langle \Phi(a_n), \Phi(a) \rangle \\ & - \epsilon \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) + \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) b_n, \end{aligned} \tag{4}$$

subject to the constraints

$$\begin{cases} \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) = 0 \\ 0 \leq \alpha_n \leq \lambda, \qquad n = 1, \ldots, N \\ 0 \leq \alpha_n^* \leq \lambda, \qquad n = 1, \ldots, N \end{cases} , \tag{5}$$

where $\langle \cdot, \cdot \rangle$ describes the dot product and α and α^* are the dual set of vari-

193 ables. The resulting SVR model is

$$f(a) = \sum_{n=1}^{N} \beta_n \langle \Phi(a_n), \Phi(a) \rangle + z \tag{6}$$

$$= \sum_{n=1}^{N} \beta_n K(a_n, a) + z, \tag{7}$$

194 where $K(a_n, a)$ is the kernel function. Any function meeting the Mercer's
195 condition can be used as the kernel function (Lu et al., 2009; Smola and
196 Scholkopf, 2004). Parameters $\beta_n = \alpha_n - \alpha_n^*$ are calculated through solving
197 the dual optimization problem and have the following relation to $\mathbf{w}$

$$\mathbf{w} = \sum_{n=1}^{N} \beta_n \Phi(a_n). \tag{8}$$

198 Since both the primal and dual optimization problem are convex, a unique
199 optimal solution can be found efficiently using numerical methods such as
200 quadratic programming (QP) (Smola and Scholkopf, 2004). Computing pa-
201 rameters β_n and z is explained in (Smola and Scholkopf, 2004) in detail.

202 In the baseline systems GMM-PCA-R and GMM-WPPCA-R (Dobry
203 et al., 2011), SVR model training and testing is implemented using LIB-
204 SVM (Chang and Lin, 2011) and the hyperparameters of the SVR such as
205 the minimal error margin ϵ and error cost factor λ are tuned using the N-fold
206 cross validation technique on the training dataset. In this research, we use
207 the same toolbox and apply the same approach to tune the hyperparameters.

208 *3.1.2. Least Squares Support Vector Regression*

209 Least Squares Support Vector Machine (LSSVM), which is a variant of
210 SVM, was introduced by Suykens and Vandewalle Suykens et al. (2002). It is

employed as a machine learning tool for classification, clustering and regression tasks. Compared to SVM, LSSVM benefits from a faster training process because the quadratic programming problem of SVM is reduced to that of solving a system of linear equations. Furthermore, the LSSVM formulation involves fewer tuning parameters (Fodor, 2003). A continuous function can be fitted to the training data with a Least Squares Support Vector Regressor (LSSVR), a technique which shares many of the advantages of LSSVM classification. In the primal form of LSSVR, which is the same as SVR, the following relation is considered for $f(a)$

$$f(a) = \mathbf{w}^t \Phi(a) + z. \tag{9}$$

In LSSVR, a least squares loss function is applied instead of Vapnik's ϵ-insensitive loss function to simplify the formulations to minimize

$$\frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{2}\gamma \sum_{n=1}^{N} e_n^2 \tag{10}$$

subject to

$$b_n = \mathbf{w}^t \Phi(a_n) + z + e_n, \tag{11}$$

where γ is a error cost factor playing the same role of λ in the SVR formulation and $e_n \in \mathbb{R}$ are error variables.

Similar to SVR, for high dimensional data this optimization problem can be solved more efficiently by introducing the Lagrangian variables ν and solving the following dual optimization problem (Suykens et al., 2002)

$$\mathbf{L}(\mathbf{w}, z, e, \nu) = \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{2}\gamma \sum_{n=1}^{N} e_n^2 \tag{12}$$

$$- \sum_{n=1}^{N} \nu_n \{\mathbf{w}^t \Phi(a_n) + z + e_n - b_n\}. \tag{13}$$

One can solve this optimization problem directly by taking the partial derivative of $\mathbf{L}$ with respect to $\mathbf{w}$, z, e and ν and setting the results to zero which leads to solving a linear system of equations. Inserting the obtained results to 9 leads to the regression function

$$f(a) = \sum_{n=1}^{N} \nu_n \langle \Phi(a_n), \Phi(a) \rangle + z \tag{14}$$

$$= \sum_{n=1}^{N} \nu_n K(a_n, a) + z, \tag{15}$$

where $K(a_n, a)$ is the kernel function and ν and z are the solution to optimization problem (12).

LSSVR has two advantages and one drawback compared to SVR. The first advantage of LSSVR is that its model training is faster as its dual form corresponds to solving a linear system which involves less computation time compared to a QP problem of SVR. The second advantage is that the LSSVR is faster to tune as its formulation involves fewer hyperparameters to tune (the minimal error margin ϵ is not used here). A drawback of this simplification is the loss of sparseness (ν is less sparse compared to β), which has been highlighted in literature (Suykens et al., 2000; Li et al., 2006).

In this research, the LSSVR models training and testing is implemented using LSSVMlab (Suykens et al., 2002) and the hyperparameters of the LSSVR are tuned on the training set using the N-fold cross validation technique.

3.2. The i-vector framework

The age estimation approaches described in section 2.1 are based on GMM mean supervectors and have been shown to yield reasonable performance. In

the related field of speaker recognition, GMM supervectors are commonplace. Recent progress in this field, however, has found an alternate method of modeling GMM supervectors that provides far superior speaker recognition performance (Dehak et al., 2011a). This technique, referred to as the i-vector framework, assumes the GMM mean supervector, $\mathbf{M}$, can be decomposed as

$$\mathbf{M} = \mathbf{u} + \mathbf{T}\mathbf{v} \tag{16}$$

where $\mathbf{u}$ is the mean supervector of the UBM, $\mathbf{T}$ spans a low-dimensional subspace (400 dimensions in this work) and subspace vector $\mathbf{v}$ is treated as a latent variable with the standard normal prior and the i-vector is its maximum-a-posteriori (MAP) point estimate.

The subspace matrix $\mathbf{T}$ is estimated via maximum likelihood in a large training dataset. An efficient procedure for training $\mathbf{T}$ and MAP adaptation of i-vectors $\mathbf{v}$ can be found in (Kenny et al., 2008). In this approach, i-vectors are the low-dimensional representation of an audio recording that can be used for classification and regression purposes.

3.3. i-vector Session Compensation

Session compensation is one of the most dominant topics in the speaker recognition field (McLaren and van Leeuwen, 2012; Dehak et al., 2011a). The main reason for using session compensation techniques is removing different session variabilities from the feature vectors (such as GMM supervectors or i-vectors) to allow the subsequent modeling approaches to better observe important between-class information. In this paper, we use Within-Class Covariance Normalization (WCCN) to normalize the within-class covariance of the i-vector space to the identity matrix (Hatch et al., 2006). In doing

so, directions of relatively high within-class variation will be attenuated and thus prevented from dominating the space (Hatch et al., 2006). The WCCN transformation matrix $\mathbf{B}_{\mathrm{W}}$ is found through Cholesky decomposition of

$$\left[\frac{1}{\jmath}\sum_{j=1}^{\jmath}\frac{1}{N_j}\sum_{i=1}^{N_j}\left(\mathbf{v}_j^i-\bar{\mathbf{v}}_j\right)\left(\mathbf{v}_j^i-\bar{\mathbf{v}}_j\right)^{'}\right]^{-1}=\mathbf{B}_{\mathrm{W}}\mathbf{B}_{\mathrm{W}}^{'}, \tag{17}$$

where $\mathbf{v}_j^i$ is the i^{th} i-vector in the j^{th} speaker, $\bar{\mathbf{v}}_j=\frac{1}{N_j}\sum_i^{N_j}\mathbf{v}_j^i$ is the mean of the observations for the j^{th} speaker, N_j denotes the number of utterances of the j^{th} speaker and $\jmath$ is the total number of speakers in the training dataset.

3.4. Train and Test

The principle of the proposed age estimation approach is illustrated in Figure 2. As it can be interpreted from this figure, in the training phase, each utterance in the training dataset is converted to an i-vector. Then, WCCN is used to remove the session variability as described in Section 3.3. Finally, the obtained vectors along with their corresponding chronological speaker age are used to train the regressor. In the testing phase, an i-vector is extracted from the utterance of an unseen speaker. Then, WCCN is used to remove the session variability. Finally, the trained regressor uses the obtained vector to estimate the chronological age of the test speaker.

The use of i-vectors for age estimation has several distinct advantages over GMM supervectors. Firstly, the relatively low dimensionality of i-vectors (400) significantly reduces the computational burden of model training and estimation compared to a GMM supervector dimensionality of greater than 12,000 used in this work. Secondly, subspace adaptation of i-vector $\mathbf{v}$ results

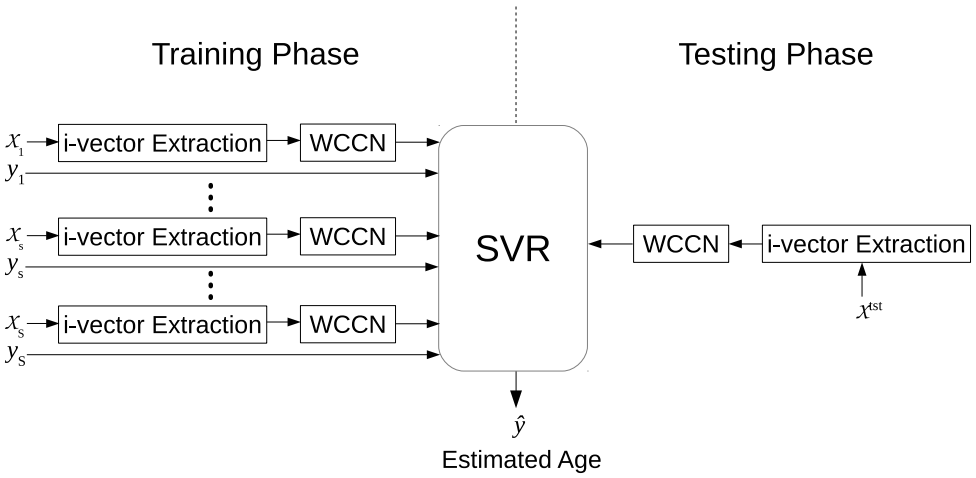

Figure 2: *The block diagram of the proposed speaker age estimation approach in training and testing phases.*

in a more reliable estimation of the true model means **M** in the context of limited training data (Dehak et al., 2011b).

4. Experimental Setup

4.1. Database

The National Institute of Standards and Technology (NIST) has held annual or biannual Speaker Recognition Evaluations (SRE) for the past two decades. With each SRE, a large database of telephone conversations (and more recently microphone speech) are released along with an evaluation protocol. These conversations typically last five minutes and originate from a large number of participants for whom meta data is recorded—including participant age and language. The NIST databases where chosen for this work due to the large number of speakers meeting the i-vector framework

requirement for a considerable amount of development data to estimate subspace matrix $\mathbf{T}$ accurately. In our experiments, first a development dataset is formed, which includes over $30,000$ speech recordings sourced from NIST 2004–2006 SRE databases, to estimate the parameters of UBM and the subspace matrix ($\mathbf{T}$). The procedure of obtaining the applied UBM and subspace matrix is presented in (McLaren and van Leeuwen, 2012).

To form the train and test datasets for speaker age estimation, telephone recordings from the common protocols of the NIST 2010 and 2008 SRE corpora are used. The core protocol, short2-short3, from the 2008 database contains 3772 telephone recordings from 1154 speakers for whom the age is between 20 and 70. The language label of 3726 utterances is given in this database. Among these, 2656 utterances are English and the remaining 1070 utterances are from 26 different non-English languages including Russian, Italian and Japanese. Similarly, the extended core-core protocol of the 2010 database contains 5479 telephone speech segments from 422 speakers for whom the age is between 20 and 70. All utterances of this database are English. There is no overlap between speech recordings extracted from the NIST 2010 and NIST 2008 SRE databases.

Figure 3 illustrates the age histograms of male and female speakers in the NIST 2010 and 2008 SRE databases. Since the perceptions of speaker gender and age have a significant mutual impact, all the experiments are performed for male and female speakers separately in this paper.

4.2. Performance Metric

The effectiveness of the applied methods is evaluated using the Mean Absolute Error (E_{ma}) of the estimated speakers' age and Pearson's correlation

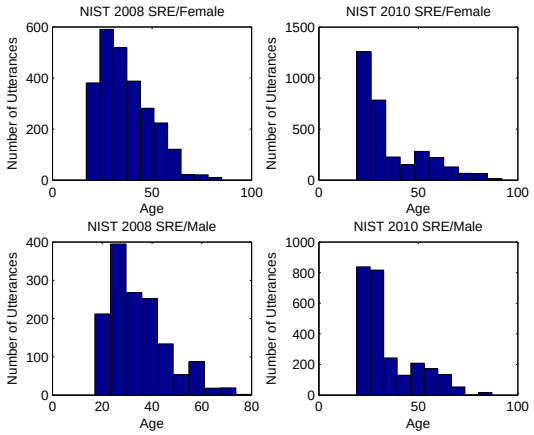

Figure 3: *Age histogram of telephone speech utterances for NIST 2010 and 2008 SRE Databases.*

coefficient (ρ) between chronological speakers' age and estimated speakers' age. The measure E_{ma} is calculated using:

$$E_{\text{ma}} = \frac{1}{Q} \sum_{q=1}^{Q} |\hat{y}_q - y_q| \,, \tag{18}$$

where $\hat{y}_q$ and y_q are the estimated and the chronological age of the q^{th} utterance of the testing dataset respectively. Q is the total number of utterances in the testing dataset. Further,

$$\rho = \frac{1}{Q-1} \sum_{q=1}^{Q} \left(\frac{\hat{y}_q - \mu_{\hat{y}}}{\sigma_{\hat{y}}} \right) \left(\frac{y_q - \mu_y}{\sigma_y} \right) , \tag{19}$$

where $\mu_{\hat{y}}$ and $\sigma_{\hat{y}}$ are the mean and the standard deviation of the speakers' estimated age respectively. Similarly μ_y and σ_y denote the mean and the standard deviation of the speakers' chronological age respectively.

We also apply the standard z-test to analyze the statistical significance level of differences between the mean absolute errors of applied systems.

5. Results and Discussion

This section presents the evaluation results of the baseline systems and compares them to the introduced i-vector based age estimation system.

The applied GMM in all experiments consist of 512 mixture components. To study the effect of the acoustic features, two types of feature vectors have been tested for the baseline systems. The first type, labeled $\text{MFCC}_{26\text{D}}$, consists of 13 Mel-Frequency Cepstrum Coefficients (MFCCs) including appended energy with their first order derivatives, forming a 26 dimensional acoustic feature vector. The second type, $\text{MFCC}_{60\text{D}}$, consists of 20 MFCCs including appended energy with their first and second order derivatives, forming a 60 dimensional acoustic feature vector. In both cases, a hamming window is used and the sampling rate, frame rate, frame size and number of Mel frequency channels are 8000 Hz, 100 Hz, 0.02 s and 30 respectively. To have more reliable features, Wiener filtering, speech activity detection (McLaren and van Leeuwen, 2011) and feature warping (Pelecanos and Sridharan, 2001) have been applied as front-end processing. The former type, $\text{MFCC}_{26\text{D}}$, matches the configuration of features applied in (Dobry et al., 2011) and the latter type, $\text{MFCC}_{60\text{D}}$, is very common in state-of-the-art i-vector based speaker recognition systems.

5.1. SVR and LSSVR

In this section, an experiment is performed to investigate the performances of SVR and LSSVR for regression in this problem and to choose the regression method with more accurate estimation results for the rest of the experiments in this paper.

In this experiment, the NIST 2008 and 2010 SRE databases are used for training and testing respectively and the acoustic features are $\text{MFCC}_{26\text{D}}$. Each utterance in the training and testing datasets is modeled using its corresponding GMM mean supervector. Then, an SVR or an LSSVR are applied as a function approximator to estimate the speakers' age.

Like the baseline systems GMM-PCA-R and GMM-WPPCA-R, SVR model training and testing is performed using LIBSVM (Chang and Lin, 2011) and the SVR Hyperparameters ϵ and λ are tuned using the 5-fold cross-validation. Since it is shown in (Dobry et al., 2011) that the radial basis function (RBF) kernel leads to more accurate estimation compared to the linear kernel, we apply the RBF kernel in our experiments. Two methods are applied to determine the width of the Gaussian functions. In the first scheme, which is adopted from (Dobry et al., 2011), the width of the Gaussian functions was set to $\sqrt{\det(\mathbf{\Sigma})/2}$, where $\mathbf{\Sigma}$ is the training feature vectors covariance matrix and $\det(.)$ denotes the determinant operator. It was mentioned in (Dobry et al., 2011) that $\sqrt{\det(\mathbf{\Sigma})/2}$ was found to be optimal on a number of empirical experiments. The results of this method, labeled as SVR 1, are listed in the first row of Table 1. In the second approach, labeled as SVR 2, the 5-fold cross-validation is used to tune the width of the Gaussian functions.

The LSSVR approach applied in this experiment also uses the RBF kernel and 5-fold cross-validation in order to tune its error cost factor and Gaussian width.

Table 1 shows the obtained results using SVR 1, SVR 2 and LSSVR in this experiment. This table shows that LSSVR estimates the speakers' age

Table 1: *The E_{ma} (in years) and ρ of male and female speakers' age estimation using SVR and LSSVR.*

Regression Method	Female		Male	
	E_{ma}	ρ	E_{ma}	ρ
SVR 1	7.59	0.80	7.97	0.69
SVR 2	7.48	0.80	7.92	0.70
LSSVR	7.44	0.80	7.87	0.70

more accurately compared to SVR 1 and SVR 2 in this experiment. LSSVR is selected for the rest of experiments in this paper rather than conventional SVR due to the obtained marginal improvement and faster and easier model training and tuning.

5.2. Baseline Systems Results

In this section, the performances of baseline systems, namely prior, GMM-R, GMM-PCA-R and GMM-WPPCA-R, are investigated.

To evaluate the baseline systems on all available utterances, 15-fold cross-validation is used. Therefore, first all speakers in the NIST 2008 and 2010 SRE databases are divided into 15 disjoint folds. Then, 15 independent experiments are run so that in each experiment, a new fold is used as the testing dataset and the remaining 14 folds are used as training dataset. Due to high variability in our data such as language, smoking habit and content, in our experiments, we have applied 15-fold rather than 5-fold or 10-fold cross-validation to have larger training datasets, which include more variability of the data.

The average E_{ma} and ρ of male and female speakers' age estimation using the baseline systems in all 15 experiments with both types of acoustic features are listed in tables 2 and 3 respectively. In this experiment, PCA and WPPCA have been tested over different target dimensions between 100 and 1000. Tables 2 and 3 only include the best results, which were obtained for target dimensions 300 and 400 for GMM-PCA-R and GMM-WPPCA-R respectively.

Results in tables 2 and 3 indicate that the GMM-R system is remarkably more accurate than the prior system. This shows that the GMM supervectors contain speaker information including age. The Tables 2 and 3 also show that the PCA and WPPCA based systems outperform the GMM-R system, thus demonstrating the benefit of dimension reduction of the GMM supervectors prior to regression. Unlike (Dobry et al., 2011) our experiments do not show remarkable advantage for using WPPCA over PCA. It is also interpreted from tables 2 and 3 that increasing the acoustic dimension from 26 to 60 slightly improves the estimation accuracy for GMM-PCA-R and GMM-WPPCA-R. Therefore, in the rest of our experiments we focused on the second type of acoustic features, MFCC_{60D}.

5.3. i-vectors for Age Estimation

The results of the proposed method for speakers' age estimation are presented in this section.

Figures 4 and 5 present the E_{ma} of the estimated age and the ρ between the chronological speakers' age and the estimated speakers' age using the proposed method and the baseline systems for different target dimensions respectively. These figures show that the proposed method, labeled i-vector-

Table 2: *The average E_{ma} (in years) of male and female speakers' age estimation for the baseline systems using* MFCC_{26D} *and* MFCC_{60D} *feature vectors.*

System Configuration	Female		Male	
	MFCC_{26D}	MFCC_{60D}	MFCC_{26D}	MFCC_{60D}
Prior	10.57	10.57	10.08	10.08
GMM-R	6.19	6.60	6.93	7.53
GMM-PCA-R	6.26	6.21	6.79	6.71
GMM-WPPCA-R	6.25	6.17	6.74	6.74

WCCN-R, is more accurate than the other state-of-the-art approaches. Note that this improvement was obtained without any optimization over the target dimension in the i-vector framework. Therefore, in figures 4 and 5, the result of proposed method is only shown for dimension 400. In the standard i-vector framework, the optimization over the target dimension is usually very time-

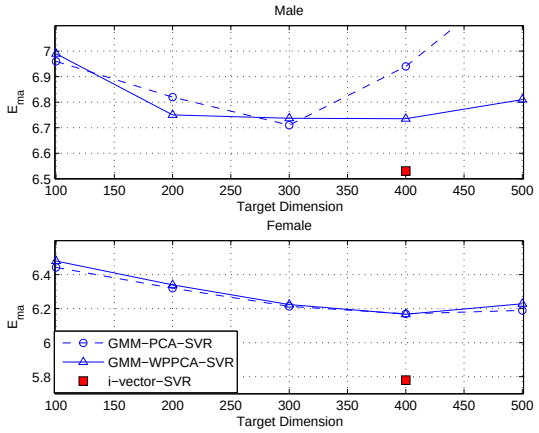

Figure 4: *The E_{ma} of female and male speakers' age estimation using the proposed method and baseline systems versus target dimension.*

Table 3: *The average ρ of male and female speakers' age estimation for the baseline systems using* $\text{MFCC}_{26\text{D}}$ *and* $\text{MFCC}_{60\text{D}}$ *feature vectors.*

System Configuration	Female		Male	
	$\text{MFCC}_{26\text{D}}$	$\text{MFCC}_{60\text{D}}$	$\text{MFCC}_{26\text{D}}$	$\text{MFCC}_{60\text{D}}$
Prior	0	0	0	0
GMM-R	0.78	0.73	0.69	0.59
GMM-PCA-R	0.77	0.78	0.71	0.72
GMM-WPPCA-R	0.77	0.78	0.71	0.71

consuming and computationally expensive. Furthermore, different studies such as (Dehak et al., 2011b) show that i-vector characteristics are mostly robust against different dimensions between 200 to 500.

The ρ and E_{ma} of age estimation using the proposed approach are 0.772 and 6.08 respectively. Therefore, the proposed method improves ρ by 12.9%,

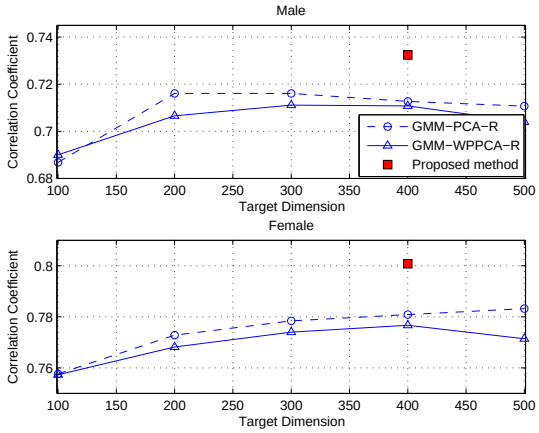

Figure 5: *Pearson correlation coefficient between estimated and true age of female and male speakers using the proposed method and baseline systems versus target dimension.*

2.0% and 2.6% relative to GMM-R, GMM-PCA-R and GMM-WPPCA-R respectively. The E_{ma} is also improved by 41%, 13%, 5% and 4.8% relative to Prior, GMM-R, GMM-PCA-R and GMM-WPPCA-R respectively. A standard z-test for comparing two means show that the E_{ma} of the i-vector based system method is significantly lower than that of the best baseline system, namely GMM-PCA-R, at the 99% confidence level. Details of this test are presented in Appendix A.

The coefficient of determination (ρ^2) (Draper and Smith, 1981) obtained by the proposed method for male and female speakers are 0.54 and 0.64 respectively, which show that roughly 54% and 64% of the variance in male and female speakers' age were successfully accounted for.

We also investigated using i-vectors without session variability compensation, like our earlier work (Bahari et al., 2012a). In this case, the ρ and E_{ma} are 0.76 and 6.22 respectively. This experiment shows that session variability compensation using WCCN relatively improves the ρ and E_{ma} by 1.5% and 2.2% respectively.

5.4. The Effect of Utterance Length

In a typical practical case, the duration of the available speech sample may vary from a few seconds to several hours. Although there is literature on the effect of available utterance duration on speaker recognition systems (Mandasari et al., 2011), there is no published research on this topic for automatic speaker age estimation systems. In this section, we analyze the performance of the proposed i-vector based speaker age estimation system with respect to speech duration in the terms of E_{ma} and ρ.

In this experiment, first all speakers in the NIST 2008 and 2010 SRE

Table 4: *The E_{ma} and ρ of speakers' age estimation using the proposed method in different test utterance length conditions.*

Utterance Length	Female		Male	
	E_{ma}	ρ	E_{ma}	ρ
5s	9.51	0.53	8.99	0.47
10s	8.5	0.64	8.27	0.57
20s	7.38	0.72	7.66	0.64
45s	6.47	0.77	6.99	0.70

databases are divided into 15 disjoint folds. Then, 15 independent experiments are run so that in each experiment, a new fold is used as testing dataset and the rest 14 folds are used as training dataset. Each utterance in the testing dataset typically contains around 80 seconds of active speech. In order to study the effect of test sample duration, we synthesized test datasets of 5, 10, 20 and 40 seconds by truncating the feature streams after speech activity detection. For consistency in our results, the test samples that contained less than 40 seconds of nominal speech using our speech detection algorithm were discarded from all results reported in this experiment. The procedure and details of obtaining corresponding i-vectors for truncated test samples is explained in (Mandasari et al., 2013). The corresponding E_{ma} and ρ values are listed in Table 4. The performance of the proposed method decreases as the test utterance duration is reduced. This is more evident when the utterance duration is less than 10 seconds. However, the results of the proposed method remain significantly more accurate than the prior, even for the utterances with a length of 5 seconds.

5.5. The Effect of Language

Braun and Cerrato performed a number of experiments to evaluate the ability of human listeners in estimating speakers' age across different languages (Braun and Cerrato, 1999). They concluded that the age can be estimated almost as accurately when the listeners are familiar with the language of the speaker as when they are not. However, Schotz considered the language as an important source influencing the acoustic analysis of speaker age (Schotz, 2006). Feld et al. studied the effect of language mismatch between train database and test samples on automatic speaker age estimation systems. In this section, we analyze the effect of language mismatch on the proposed i-vector based age estimation system.

In this experiment, the train database is NIST 2010 SRE, which includes 5634 English utterances from 445 speakers. There are two test databases in this experiment, the English and non-English parts of the NIST 2008 SRE database. Figure 6 illustrates the age histograms of the English and non-English speakers of the NIST 2008 SRE database. To eliminate the effect of utterance length, we synthesized test samples of 40 seconds by truncating the feature streams after speech activity detection. The E_{ma} and ρ of this experiment for both English and non-English test sets are listed in table 5.

Table 5: *The E_{ma} and ρ for both English and non-English test sets.*

System Configuration	Female		Male	
	English	Non-English	English	Non-English
E_{ma}	6.92	8	7.72	8.32
ρ	0.66	0.42	0.50	0.32

Results in table 5 indicate that language mismatch between train database and test samples causes a large performance degradation in both E_{ma} and ρ. It is obvious that the E_{ma} for the English test set is significantly less than that of the non-English test set for both male and female utterances. In these experiments, since only telephone speech signals are used, we do not concentrate on channel mismatch. The effect of gender is also discarded because all the experiments are performed for male and female speakers separately.

6. Conclusions

In this paper, utterance modeling with i-vectors, which was successfully applied to speaker recognition, has been used in conjunction with a WCCN and a LSSVR to address speaker age estimation. For the evaluation, telephone utterances of NIST 2010 and 2008 SRE databases have been used.

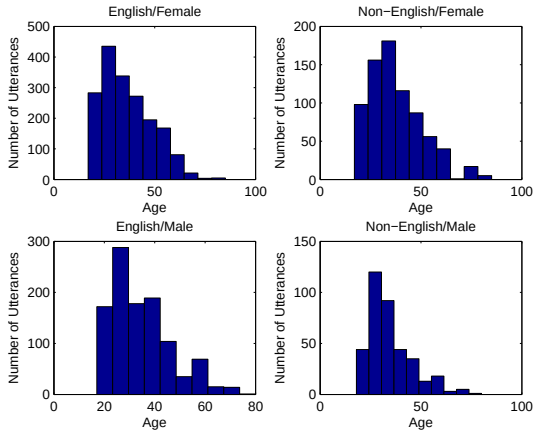

Figure 6: *Age histogram of English and non-English speakers in the NIST* 2008 *SRE database.*

Assessment results demonstrate that ρ and E_{ma} for the proposed approach are 0.772 and 6.08 respectively. Therefore, the obtained relative improvements of ρ and E_{ma} compared to the best baseline system are around 2% and 5% respectively. The experiments on analyzing the effect of utterance duration reveals that the performance of the proposed method degrades as the utterance length decreases especially for samples shorter than 20 seconds. However, it is still more accurate than the prior baseline system even for utterances of 5 seconds in length. Analyzing the effect of language shows that the language mismatch between train and test databases significantly decreases the performance of the age estimation system.

7. Acknowledgements

This work is supported by the European Commission as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008), namely Bayesian Biometrics for Forensics (BBfor2), under Grant Agreement number 238803.

Appendix A.

In this appendix, a statistical analysis is presented to compare the mean absolute errors of age estimation obtained by the i-vector-SVR and GMM-PCA-R.

Since the values of populations variances are unknown, tests for the comparison of two means should be conducted with the a t-test normally. However, both sample sizes are greater than 30 in this case and we can work with the standard normal distribution (z-test) instead of Student distribution (t-test). In the standard z-test for comparison of two means, the z value is

calculated as follows:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{A.1}$$

where $\bar{x}_i$, s_i, and n_i denote the mean, the variance and total number of samples in the first set respectively. Similarly, $\bar{x}_2$, s_2, and n_2 are the mean, the variance and sample size in the second set respectively.

In the comparison of the mean absolute errors of age estimation obtained by the i-vector-SVR ($\bar{x}_1$) and GMM-PCA-R ($\bar{x}_2$), the null hypothesis is $\bar{x}_2 \leq \bar{x}_1$ and the alternative hypothesis is $\bar{x}_2 > \bar{x}_1$. With significance levels $\alpha = 0.01$ and $\alpha = 0.05$, the critical values of z are 2.33 and 1.645 respectively for a one tail test.

The mean and the standard deviation of the age estimation absolute error using i-vector-SVR and GMM-PCA-R over male and female utterances are listed in Table A.6.

As it is shown in Table A.6, the obtained z for male and female utterances is greater than critical value of z for significance levels $\alpha = 0.05$ and $\alpha = 0.01$ respectively. Therefore, the null hypothesis is rejected and it is concluded that the alternative hypothesis is true.

In the test of significance, we are trying to compare GMM-WPPCA-R and the proposed method. Consequently, all results of the proposed method (regardless of gender) can be considered in one class and all the results of GMM-WPPCA-R are assumed to be in the other class. The last row of Table A.6 shows the mean and the standard deviation of the age estimation absolute error using the proposed method and GMM-WPPCA-R over all utterances regardless of gender (labeled both). The obtained z value of this experiment is 4.15 which is greater than the critical value of z for significance

Table A.6: *The mean and the standard deviation of age estimation absolute error using i-vector-SVR and GMM-PCA-R over male and female utterances.*

Gender	Parameter	Proposed method	GMM-WPPCA-R	z
Male	$\bar{x}_i$	6.53	6.74	1.7
	s_i	5.36	5.54	
	n_i	3883	3883	
Female	$\bar{x}_i$	5.78	6.17	4.13
	s_i	4.78	4.92	
	n_i	5292	5292	
Both	$\bar{x}_i$	6.10	6.41	4.15
	s_i	5.05	5.20	
	n_i	9175	9175	

level $\alpha = 0.01$.

References

Bahari, M.H., McLaren, M., Van Leeuwen, D., Van hamme, H., 2012a. Age estimation from telephone speech using i-vectors, in: Proc. Interspeech, pp. 506–509.

Bahari, M.H., Saeidi, R., Van hamme, H., van Leeuwen, D., 2013. Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech, in: International conference on acoustics, speech, and signal processing, pp. 7344–7348.

570 Bahari, M.H., Van hamme, H., 2011. Speaker age estimation and gender
571 detection based on supervised non-negative matrix factorization, in: Proc.
572 IEEE Workshop on Biometric Measurements and Systems for Security and
573 Medical Applications (BIOMS), pp. 1–6.

574 Bahari, M.H., et al., 2012b. Speaker age estimation using hidden markov
575 model weight supervectors, in: 11th IEEE Int. Conf. Information Science,
576 Signal Processing and their Applications (ISSPA), pp. 517–521.

577 Bocklet, T., Maier, A., Bauer, J., Burkhardt, F., Noth, E., 2008a. Age and
578 gender recognition for telephone applications based on GMM supervectors
579 and support vector machines, in: Proc. IEEE Int. Conf. Acoustics, Speech
580 and Signal Processing (ICASSP), pp. 1605–1608.

581 Bocklet, T., Maier, A., Noth, E., 2008b. Age determination of children
582 in preschool and primary school age with GMM-based supervectors and
583 support vector machines-regression, in: Proc. Text, Speech and Dialogue,
584 pp. 253–260.

585 Bocklet, T., Stemmer, G., Zeissler, V., Noth, E., 2010. Age and gender
586 recognition based on multiple systems early vs. late fusion, in: Proc. 11th
587 Annual Conference of the International Speech Communication Associa-
588 tion, pp. 2830–2833.

589 Braun, A., Cerrato, L., 1999. Estimating speaker age across languages, in:
590 Proc. ICPhS, pp. 1369–1372.

591 Campbell, W., Sturim, D., Reynolds, D., 2006. Support vector machines

using GMM supervectors for speaker verification. IEEE Signal Processing Letters 13, 308–311.

Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 27.

Chen, C.C., Lu, P.T., Hsia, M.L., Ke, J.Y., Chen, O.C., 2011. Gender-to-age hierarchical recognition for speech, in: Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on, pp. 1–4.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech, and Lang. Process. 19, 788–798.

Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D., Dehak, R., 2011b. Language recognition via ivectors and dimensionality reduction, in: Proc. Interspeech, pp. 857–860.

Dobry, G., Hecht, R., Avigal, M., Zigel, Y., 2009. Dimension reduction approaches for svm based speaker age estimation, in: Proc. Interspeech, pp. 2031–2034.

Dobry, G., Hecht, R.M., Avigal, M., Zigel, Y., 2011. Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. IEEE Trans. Audio, Speech, and Lang. Process. 19, 1975–1985.

Draper, N.R., Smith, H., 1981. Applied regression analysis 2nd ed. .

Feld, M., Barnard, E., van Heerden, C., Muller, C., 2009. Multilingual speaker age recognition: Regression analyses on the lwazi corpus, in: Au-

tomatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, pp. 534–539.

Fodor, I., 2003. Statistical techniques to find similar objects in images, in: Proc. the American Statistical Association, Statistical Computing Section.

Hatch, A., Kajarekar, S., Stolcke, A., 2006. Within-class covariance normalization for svm-based speaker recognition, in: Proc. Interspeech.

van Heerden, C., Barnard, E., Davel, M., van der Walt, C., van Dyk, E., Feld, M., Muller, C., 2010. Combining regression and classification methods for improving automatic speaker age recognition, in: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp. 5174–5177.

Kelly, F., Drygajlo, A., Harte, N., 2013. Speaker verification in score-ageing-quality classification space. Computer Speech & Language .

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of interspeaker variability in speaker verification. IEEE Trans. Audio, Speech, and Lang. Process. 16, 980–988.

Kockmann, M., Burget, L., Cernocky, J., 2010. Brno university of technology system for interspeech 2010, in: Proc. 11th Annual Conference of the International Speech Communication Association, pp. 2822–2825.

van Leeuwen, D.A., Bahari, M.H., 2012. Calibration of probabilistic age recognition, in: Proc. Interspeech, pp. 502–505.

Li, M., Han, K.J., Narayanan, S., 2013. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. Computer Speech and Language 27, 151 – 167.

Li, Y., Lin, C., Zhang, W., 2006. Improved sparse least-squares support vector machine classifiers. Neurocomputing 69, 1655–1658.

Linville, S.E., 2001. Vocal aging. Singular Thomson Learning.

Lu, C., Lee, T., Chiu, C., 2009. Financial time series forecasting using independent component analysis and support vector regression. Decision Support Systems 47, 115–125.

Mahmoodi, D., Soleimani, A., Marvi, H., Razzazi, F., Taghizadeh, M., Mahmoodi, M., 2011. Age estimation based on speech features and support vector machine, in: 3rd Computer Science and Electronic Engineering Conference, pp. 60–64.

Mandasari, M., McLaren, M., van Leeuwen, D., 2011. Evaluation of i-vector speaker recognition systems for forensic application, in: Proc. Interspeech, pp. 21–24.

Mandasari, M.I., Saeidi, R., McLaren, M., van Leeuwen, D.A., 2013. Quality measure functions for calibration of speaker recognition system in various duration conditions. IEEE Transactions on Acoustics, Speech, and Language Processing 21, 2425 – 2438.

McLaren, M., van Leeuwen, D., 2011. A simple and effective speech activity detection algorithm for telephone and microphone speech. Proc. NIST SRE Workshop , 1–6.

McLaren, M., van Leeuwen, D., 2012. Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources. IEEE Trans. Audio, Speech, and Lang. Process. 20, 755–766.

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J., et al., 2007. Comparison of four approaches to age and gender recognition for telephone applications, in: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. IV–1089.

Minematsu, N., Sekiguchi, M., Hirose, K., 2002. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers, in: IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), pp. I–137.

Muller, C., Wittig, F., Baus, J., 2003. Exploiting speech for recognizing elderly users to respond to their special needs, in: Proc. 8th European Conf. Speech Communication and Technology (Eurospeech), pp. 1305–1308.

Mysak, E.D., 1959. Pitch and duration characteristics of older males. Journal of Speech, Language and Hearing Research 2, 46.

Nguyen, P., Le, T., Tran, D., Huang, X., Sharma, D., 2010. Fuzzy support vector machines for age and gender classification, in: Proc. 11th Annual Conference of the International Speech Communication Association, pp. 2806–2809.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification , 213–218.

Porat, R., Lange, D., Zigel, Y., 2010. Age recognition based on speech signals using weights supervector, in: Proc. 11th Annual Conference of the International Speech Communication Association, pp. 2814–2817.

Schotz, S., 2006. Perception, analysis and synthesis of speaker age. volume 47. Citeseer.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Muller, C., Narayanan, S., 2010. The interspeech 2010 paralinguistic challenge, in: Proc. Interspeech, pp. 2794–2797.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., Narayanan, S., 2013. Paralinguistics in speech and languagestate-of-the-art and the challenge. Computer Speech & Language 27, 4–39.

Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures, in: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 31–36.

Smola, A., Scholkopf, B., 2004. A tutorial on support vector regression. Statistics and computing 14, 199–222.

Suykens, J.A., Lukas, L., Vandewalle, J., 2000. Sparse approximation using least squares support vector machines, in: Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on, pp. 757–760.

Suykens, J.A.K., et al., 2002. Least squares support vector machines. World Scientific.

Tanner, D.C., Tanner, M.E., 2004. Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection. Lawyers and Judges Publishing.